INFLUENCE OF SOCIAL MOTIVATION OVER BELIEF DYNAMICS: A GAME-THEORETICAL ANALYSIS

Fabio Paglieri[1], Cristiano Castelfranchi[2]
[1]Department of Philosophy and Social Sciences, University of Siena
paglieri@media.unisi.it
[2]Institute for Cognitive Sciences and Technologies, National Research Council (ISTC-CNR)
cristiano.castelfranchi@istc.cnr.it

ABSTRACT

This paper provides a game-theoretical description of social and motivational influence over belief dynamics of two arguing agents that hold contrasting views. The formal analysis shows how social influence depends on both (1) the agent's own motives and (2) her beliefs concerning the motives of the other agent. Moreover, game-theoretical modelling of dialogical interaction with mutual ignorance of the agents' motivational profiles reveals that (3) some information on the counterpart's motivations have greater diagnostic value than others, and (4) such discontinuity in informational value again depends on the agent's own motivations. Hence this game-theoretical analysis captures several features of the rich interplay between motivations and beliefs, and it allows for prediction of the specific motivational pressure being exerted on the agent's belief dynamics, given her current frame of mind. Here the model is applied both to agents with exhaustive knowledge of each other preferences, and to agents with only partial assumptions on the motivational profile of their counterpart. In the final sections, we discuss the extension of this analysis to the single-agent case, and future empirical verification of the predictions generated by the model, via social experiments (e.g. experimental economics) and computational models (e.g. agent-based social simulation).

INTRODUCTION

Social influence over belief formation and change had been theoretically modelled and empirically verified both in social psychology (Asch, 1952; Kruglanski, 1980; Kunda, 1990; Castelfranchi & Miceli, 1998; Forgas, 2000) and in philosophical epistemology (Goldman, 1999). These studies suggest that beliefs do not simply capture a relation between an agent and the world, as the classical notion of *episteme* as 'true belief' implies. Beliefs are also *social entities*, depending on social relations for their origins and support (they often derive from social sources and are justified by the fact that others believe so; Asch, 1952), for their *use* (e.g. communication, coordination), and for their *functions* (e.g. to be shared, to provide a common ground).

In contrast with these results, standard formalisms of belief dynamics such as AGM-style belief revision (Alchourrón et al., 1985; Gärdenfors, 1988), Truth-Maintenance Systems (Huns & Bridgeland, 1991; Doyle, 1992) and probabilistic models (Berger, 1985; Boutilier, 1998) usually fail to consider motivation, both social and individual, as a relevant factor in defining and shaping the agent's belief set. While the *coordination* of motivation and beliefs is a typical problem in agent architectures (Castelfranchi, 1998), the *influence* of motivation over belief formation and change is not usually investigated. The same applies to Bayesian analyses: although the connection between decision-making and beliefs is an obvious focus of interest (Berger, 1985), these approaches typically keep utility (motivation) quite separate from probability (belief), without addressing the influence of the former on the latter. As for AGM belief revision, it might be said to be implicitly driven by the 'motivation' of maintaining a coherent set of beliefs, avoiding contradictions when faced with new information in contrast with previous convictions (Harman, 1986; Gärdenfors, 1988; Levi, 1991). However, this is rather a basic assumption of the model, which does not play a specific role in orienting the agent toward believing or disbelieving any particular claim.

This lack of interest for social and motivational influence in formal models of belief dynamics is quite puzzling, if contrasted with the overwhelming evidence of such influence provided by both empirical and theoretical research in cognitive psychology (Festinger, 1957; Kruglanski, 1980; Swann, 1990; Castelfranchi, 1996; Paglieri, 2004), social psychology (Asch, 1952; Kunda, 1990; Miceli & Castelfranchi, 1998; Forgas, 2000), economics (Kahneman & Tversky, 1979) and computer science (Picard, 1997). In natural cognitive agents (e.g. humans), motivation does play a role in shaping the agent's beliefs, both by focusing her attention on those issues and data she considers more relevant and urgent (Kruglanski, 1980; Kunda, 1990), and by orienting her assessment of available information according to specific patterns, such as social conformity (Asch, 1952), confirmatory tendency (Festinger, 1957), self-verification (Swann, 1990), denial (Miceli and Castelfranchi, 1998). Moreover, similar motivational biases are not necessarily irrational or anti-adaptive: while relevance-based belief assessment can lead to tunnel vision and even obsession, some kind of goal-directed focusing is indeed needed for any resource-bounded agent, in order to efficiently perform belief revision (Cherniak, 1986); confirmation and self-verification can degenerate in wishful thinking and self-delusion, but they are also basic defence mechanisms to preserve from loss of motivation, and they often serve as simple effective heuristics for specific tasks (Todd & Gigerenzer, 2000).

However, the aim of this paper is not to make the case for motivational influence over belief dynamics in general, but rather to focus on *social motivation* within a specific formal framework, i.e. game theory. The basic research questions addressed here are the following: Provided social motivation does play a role in belief dynamics, what *exactly* is this role? Is it always one and the same, or does it depend on the *motivational profile* of the agent, i.e. different motivations might affect in different ways the agent's belief dynamics? Is there any way of *predicting the specific social impact over belief change*, given the agent's motivations?

Our work tackles these questions by applying a game-theoretical approach to belief change in a *dialogical setting*, in which two agents (called P and ~P) debate a controversial point on which they have mutually excluding views, and each of them has to decide whether to maintain her own view or change it and assume the view of the opponent. Different motivational profiles (i.e. preference orderings over possible outcomes) are defined and contrasted, highlighting the emergent *belief revision strategies* of the agents – with either complete or partial assumptions on each other preferences. The formal analysis emphasizes *social influence* in shaping individual beliefs, and the pivotal role played by motivation in this process.

MOTIVATED DIALOGUES: TWO AGENTS ARGUING WITH EACH OTHER

Imagine two agents confronting each other on a given issue, on which they hold contrasting and mutually excluding views. Let us call these agents P and ~P, simply to signify their initial disagreement on the issue under consideration. Each agent is faced with an alternative: either she maintains (M) her own view, or she revises (R) it in favor of the opponent's claim. This produces four possible outcomes.

		~P	
		Maintain	**Revise**
P	**Maintain**	*disagreement (conservative)*	*agreement P*
	Revise	*agreement ~P*	*disagreement (role-reversal)*

Table1. Possible outcomes of interaction.

Whenever one agent revises her previous claim, while the other maintains her own view, the dialogue ends with a mutual agreement – either on P or ~P. Otherwise, the initial disagreement still holds, either because of a general conservative attitude (both agents maintain their own views), or because of a complete role-reversal, in which both agents revise their claims, so they end up to be again on opposite sides of the barricade – they just traded places.

This dialogical setting is obviously highly idealized, since here we assume an all-or-nothing situation in which *tertium non datur*, i.e. the agent cannot withdraw her previous claim without converging on the opponent's view. In real-life argumentation, this is not always the case. Imagine for instance two agents debating on whether option A is better than B, or vice versa. The first agent holds that A is better than B, while the second agent believes B to be better than A: in this case, they might come to an agreement by dropping both their claims, and concluding that options A and B are simply indifferent. More generally, people usually do not argue on isolated claims, but rather on complex issues with several features. One of the most viable ways towards agreement is *compromise*, that can be roughly understood as the practice of conceding to the opponents on some points, while maintaining one's own view on other aspects. If we include compromise as a possible move for our arguing agents, we obtain a richer picture of the possible interactive outcomes. Or we can even go a step further, characterizing compromise itself as a continuum, as showed on the lower half of Figure 1.

		~P		
		Maintain	**Compromise**	**Revise**
	Maintain	*disagree (conservative)*	*disagree*	*agree P*
P	**Compromise**	*disagree*	*agreement (compromise)*	*disagree*
	Revise	*agree ~P*	*disagree*	*disagree (reversal)*

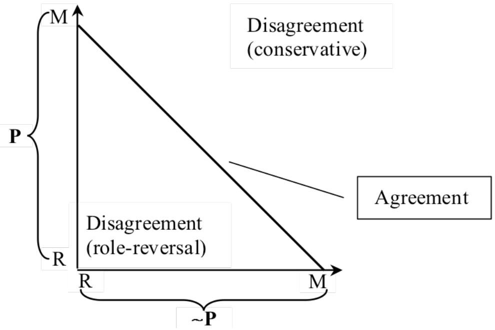

Figure 1. Modelling compromise: Discrete vs. continuous analysis

However, in this paper we will confine ourselves to the most straightforward case of dialogical interaction, the all-or-nothing situation described in Table 1. This will serve to make our analysis clear and easy to follow, even from readers not familiar with game-theoretical modeling. Besides, the extension of our methods and results to more complex cases of interaction, as those summarized in Figure 1, is a mere matter of technical exercise, and it does not imply any significant divergence from the basic principles discussed here.

So, given this simple dialogical setting, the question is: What would an agent *prefer* to do – maintain or revise her view? Under which conditions each option would be better, and why?

Approaches which deny or minimize the role of motivation in belief dynamics would answer that (1) it does not matter the option chosen by the opponent in determining the preferred option of the agent, since belief formation should rest only on epistemic factors, such as factual credibility and hard evidence; for the same reason, (2) it does not even matter what are the agent's own preferences concerning social outcomes of the dialogue (agreement or disagreement), because she should not be biased by her will in assessing her beliefs.

While non-negotiability of beliefs (i.e., the fact that an agent is not free to believe at will, but she rather needs convincing reasons to do so) is indeed a distinctive feature of cognitive systems[1] (Harman, 1986; Levi, 1991; Castelfranchi, 1996), at the same time motivational influences are equally well-known hallmarks of belief dynamics (Festinger, 1957; Kunda, 1990; Paglieri, 2004): hence «[t]he biasing role of goals is thus constrained by one's ability to construct a justification for the desired conclusion: people will come to believe what they want to believe only to the extent that reason permits» (Kunda, 1990: 483). With reference to our dialogical setting, the point is to see whether is possible to define a specific outcome such as to be the agent's 'desired conclusion', once given the agent's own goals (here represented in terms of preference orderings) and her beliefs concerning the goals of her opponent. This desired conclusion will *not* represent the deterministic outcome of process (what the agent will actually believe), but rather *the focus of the motivational pressure applied over her belief dynamics* (what she would like to believe). As Kunda remarked, the resulting influence and bias will necessarily be only indirect, e.g. in terms of selective attention, reasoning patterns, choice of heuristics, etc. Here our purpose is not to detail the nature of such influence, but rather to frame it in a strategic setting and *to predict, given the individual motivations of the two agents, their rational preferences concerning the social outcomes*.

MOTIVATIONAL PROFILES IN DIALOGICAL BELIEF DYNAMICS

To model realistic motivational profiles on belief dynamics for cognitive agents in a dialogical setting, we first define the preferences of each agent on both the outcome of her own belief dynamics (*cognitive attitudes*), and the integration with the other agent's belief dynamics (*social attitudes*). Technically, these basic attitudes are modeled as *partial orderings over possible outcomes* of the dialogue.

Attitude	**Criterion (partial ordering)**
Cognitive:	
CONSERVATIVE	Maintain is better than Revise *(MM or MR) > (RM or RR)*
EXPLORATIVE	Revise is better than Maintain *(RM or RR) > (MM or MR)*
Social:	
COOPERATIVE	Agree is better than Disagree *(MR or RM) > (MM or RR)*
ANTAGONISTIC	Disagree is better than Agree *(MM or RR) > (MR or RM)*

Table 2. Basic attitudes in belief dynamics.

Each agent is characterized by both cognitive and social attitudes, i.e. preferences on the outcome that involve both her own cognitive processes and their integration with those of the counterpart. The point is, which one of those attitude is dominant in the agent's mind? Is the agent either *individual-oriented* (i.e. the cognitive attitude rules over the social attitude), or is she *social-oriented* (the social attitude is prominent with respect to the cognitive attitude)? By answering this question, we are capable of capturing eight different *motivational profiles* (i.e. total preference orderings over the possible outcomes of the dialogue) – four of them individual-oriented, and four social-oriented. Each of these motivational profile roughly

[1] Philosophically oriented readers might want to compare non-negotiability of beliefs with *doxastic voluntarism*, as a partially opposed view of belief formation (see Wansing, 2000 for a recent survey). However, since this debate does not bring any major consequence on the present study, we will skip further references to doxastic voluntarism in this paper.

describes a possible 'psychological type' or 'personality', and it is liable to influence the agent's belief dynamics in different ways.

Individual-oriented Motivational Profiles

		Primary Cognitive Attitude	
		CONS	EXPL
Secondary Social Attitude	COOP	*Despotic* *(MR>MM>RM>RR)*	*Pliable* *(RM>RR>MR>MM)*
	ANTA	*Headstrong* *(MM>MR>RR>RM)*	*Whimsical* *(RR>RM>MM>MR)*

Social-oriented Motivational Profiles

		Primary Social Attitude	
		COOP	ANTA
Secondary Cognitive Attitude	CONS	*Open-minded* *(MR>RM>MM>RR)*	*Contentious* *(MM>RR>MR>RM)*
	EXPL	*Agreeable* *(RM>MR>RR>MM)*	*Polemist* *(RR>MM>RM>MR)*

Table 3. Motivational profiles in dialogical belief dynamics

Here a certain motivational profile is understood as emerging from the integration of both cognitive and social attitudes of the agent, depending on (1) the nature of such attitudes[2], and (2) their relative importance for the agent. For instance, an agent that is both conservative and cooperative, but that gives priority to conservationism over cooperation, is stigmatized as a *Despotic* agent, since she tries to reach an agreement only on her own views, and is willing to give up the agreement with the counterpart, if this would force her to change her own opinions. In contrast, an *Open-minded* agent is once again both conservative and cooperative, but here the latter tendency is stronger than the former: hence the agent, although preferring as optimal an agreement on her own terms, is ready to drop her previous claim, rather than jeopardizing cooperation. Clearly enough, despotism is a typical individual-oriented attitude (the private outcome is more relevant than the social one), while an open-mind motivational profile is strongly social-oriented (the social outcome has priority over the agents' own preferences over their individual beliefs). Similar considerations apply to all the motivational profiles summarized in Table 3. Which brings us to the crucial point of this analysis: Given a specific preferences profile for each agent, what kind of influence (motivational and social) is to be expected over belief dynamics in their interaction?

SOCIAL INFLUENCE OVER BELIEF DYNAMICS WITH KNOWLEDGE OF THE OPPONENT'S MOTIVATIONS

In this section we assume that both agents know exactly (1) their own preferences over possible outcomes, and (2) the motivational profile of each other; on the basis of such assumptions, we are able to work out their desirable outcome and preferred option. Notice that here, as in the rest of this paper, 'desirable outcome' and 'preferred option' are meant as technical means to model motivational and social pressures: hence the 'desirable outcome' is not necessarily the solution that the agent will achieve or even try to achieve, but rather the state of the world that would be most pleasant to reach; similarly, her 'preferred option' simply captures the direction (either to maintain or to revise old convictions) towards which the agent's belief dynamics *might* be in fact biased.

This said, let us consider first the case with two players who share the same motivational profile. Perhaps not surprisingly, it turns out that for all pairs of individual-oriented types there is only one Nash equilibrium, hence the dominant pressure is straightforward – as exemplified in Table 4 with two Whimsical agents, that would be both happy to change their initial views and fail to achieve mutual agreement (in tables from now

[2] For the sake of simplicity, here we consider only the cases where the agent has a total preferences ordering over the outcomes of the dialogical interaction, i.e. no outcome is considered equally preferred or indifferent with respect to any other. We leave the treatment of interaction of agents with incomplete motivational profiles (i.e. partial orderings over the outcomes) to future works.

on, the first pay-off refers to the row-player, the second to the column-player).

		~P	
		Maintain	**Revise**
P	**Maintain**	*1, 1*	*0, 2*
	Revise	*2, 0*	***3, 3***

Table 4. Interaction of Whimsical agents

In contrast, whenever a social-oriented type meets her analogous, there are always two Nash equilibria, so that the game needs to be solved in its extended form[3], i.e. representing the development of the interaction over time – implying that the agent who moves first in fact chooses between two final outcomes, since she can effectively predict the next move of her counterpart[4]. Table 5 shows the game-theoretical representation of the interaction of two Contentious agents, both in strategic and extended form: here the first to move is actually choosing between the alternative outcomes MM and RR, and obviously her preferred strategy is to choose the best one for her (MM).

This frame of analysis does not change when the interaction is between players with different motivational profiles, e.g. when P is Despotic and ~P is Agreeable: while the pay-off are obviously different and the matrix is no more symmetric, the situation is still analogous, and all these games are easily solved, either in strategic or extended form (Table 6 summarized the Despotic-Agreeable case).

		~P	
		Maintain	**Revise**
P	**Maintain**	***3, 3***	*1, 0*
	Revise	*0, 1*	***2, 2***

Solved in extended form as follows:

1st player	2nd player	Outcome
M	**M**	***3, 3***
	R	*1, 0*
R	**M**	*0, 1*
	R	*2, 2*

Table 5. Interaction of Contentious agents

		Agreeable ~P	
		M	**R**
Despotic P	**M**	*2, 0*	***3, 3***
	R	*1, 2*	*0, 1*

Table 6. Interaction of Despotic P with Agreeable ~P

Of course, different pairs of attitudes would lead to different preferred outcomes: e.g., while agreement with a Despotic P is in the best interest of an Agreeable ~P, the same does not apply if ~P is instead a Polemist – in which case the emergent preferred outcome would be a conservative disagreement, i.e. each agent would rather prefer to keep her own views on the matter.

More interesting are the cases with *two social-oriented agents with different motivational attitudes*. As we noticed before, these cases require to represent the game in extended form and to assess a sub-game perfect equilibrium through backward induction. However, since now the agents do not share the same motivational profile, such equilibrium will change depending on which

[3] Interaction of social-oriented agents needs a more sophisticated analysis exactly because they are inclined to *coordinate with others via integration of their cognitive attitudes*.

[4] Technically speaking, such preferred outcome is a *sub-game perfect equilibrium* derived via *backward induction* – moreover, it is the only sub-game perfect equilibrium available in this interaction. Interested readers may easily verify that, for any interaction between agents who share the same social-oriented attitude, there is one and only one sub-game perfect equilibrium, i.e. one and only one preferred outcome, assuming precise knowledge of the counterpart's preferences.

agent moves first[5] in the interaction. This can be illustrated by considering the interaction between an Open-minded P and a Polemist ~P (Table 7): here the sub-game perfect equilibrium is MM whenever the Open-minded agent moves first, while it is RM when it is the Polemist to make the initial move (notice that the pay-offs in Table 7 are ordered such as to indicate first the pay-off of the agent who initiates the game, and then that of the other agent).

When Open-minded moves first:

Open-minded	**M**		**R**	
Polemist	**M**	**R**	**M**	**R**
Outcome	***1, 2***	*3, 1*	*2, 0*	*0, 3*

When Polemist moves first:

Polemist	**M**		**R**	
Open-minded	**M**	**R**	**M**	**R**
Outcome	*2, 1*	*0, 2*	***1, 3***	*3, 0*

Table 7. Interaction of Open-minded P with Polemist ~P

Applying the same line of reasoning to all possible combinations of motivational profiles, we are able to derive what is the preferred outcome for each of them, given the agent's assumptions on the preferences of the other player (as summarized in Table 8).

			Individual				**Social**			
			Cons		Expl		Coop		Anta	
			Ds	Hd	Pl	Wh	Op	Ag	Ct	Po
Ind	Cons	Ds	*MM*	*MM*	*MR*	*MR*	*MR*	*MR*	*MM*	*MM*
		Hd	*MM*	*MM*	*MR*	*MR*	*MR*	*MR*	*MM*	*MM*
	Expl	Pl	*RM*	*RM*	*RR*	*RR*	*RM*	*RM*	*RR*	*RR*
		Wh	*RM*	*RM*	*RR*	*RR*	*RM*	*RM*	*RR*	*RR*
Soc	Coop	Op	*RM*	*RM*	*MR*	*MR*	*MR*	*MR*	*MM*	*MM*
		Ag	*RM*	*RM*	*MR*	*MR*	*RM*	*RM*	*RR*	*RR*
	Anta	Ct	*MM*	*MM*	*RR*	*RR*	*MR*	*MR*	*MM*	*MM*
		Po	*MM*	*MM*	*RR*	*RR*	*RM*	*RM*	*RR*	*RR*

Table 8: Preferred outcomes under different assumptions on the other agent's motivations

SOCIAL INFLUENCE OVER BELIEF DYNAMICS UNDER IGNORANCE OF THE OPPONENT'S MOTIVATIONS

So far we worked under the assumption that both agents know exactly the preferences of each other – or better, they are subjectively sure of such preferences[6]. But this does not need to be the case between real agents, both natural and artificial, that, more often than not, will interact without having either precise or exhaustive knowledge of each other motivational profiles. Hence we need to relax this constraint, if we want (as it is the case here) to apply this formal analysis to realistic cognitive agents[7].

[5] It is important to clarify what does it means, in the context of this analysis, 'to move first', since here we are *not* using game theory to predict the actual behaviour of the agents (because what we believe depends on our preferences *only to a limited extent*), but rather to predict *the specific influence that is to be expected* over the agent's belief dynamics, given her motivational profile and her assumptions on the motivational profile of the counterpart. Under these conditions, the agent who moves first simply represents *the agent whose perspective we are considering* (as observers) in assessing motivational influence over belief dynamics. For instance, if we want to capture motivational pressures over belief dynamics of an Open-minded agent who assumes to face a Polemist opponent, then our analysis predicts that the agent will be biased towards maintaining her previous belief and making this move explicit, knowing that a conservative disagreement will then be the preferred outcome for the other agent (see Table 7).

[6] To shape motivational influence over belief dynamics, it does not matter that the agent's convictions on the opponent's preferences are objectively correct (i.e. true): subjective certainty is enough.

[7] This move is not new in game theory: e.g., *Variable Frame Theory* (VFT) deals exactly with similar issues (see Bacharach, 1993 for a formal

This raises an interesting challenge to our model: Provided that (1) motivational pressures depend also on the agent's beliefs on the opponent's preferences, but (2) she has no sure insight on such preferences, how can we assess the motivational influence (if any) on her belief dynamics?

To answer this question, we need to look again at the results summarized in Table 8, with the aim of discriminating between *useful* and *needless* information on the opponent's preferences – that is, we check under which conditions the preferred outcome for an agent with a given motivational profile is *unaffected* by the (presumed) motivational profile of the counterpart. We first start by clustering together those motivational profiles with identical sets of preferred outcomes, as showed in Table 9.

		Assumption on the counterpart					
		Cons	Expl	Open	Agree	Cont	Pole
Agent's profile	Cons	*MM*	*MR*	*MR*	*MR*	*MM*	*MM*
	Expl	*RM*	*RR*	*RM*	*RM*	*RR*	*RR*
	Open	*RM*	*MR*	*MR*	*MR*	*MM*	*MM*
	Agre	*RM*	*MR*	*RM*	*RM*	*RR*	*RR*
	Cont	*MM*	*RR*	*MR*	*MR*	*MM*	*MM*
	Pole	*MM*	*RR*	*RM*	*RM*	*RR*	*RR*

Table 9. Simplified matrix of preferred outcomes

Table 9 highlights that discriminating between Despotic and Headstrong opponents, or between Pliable and Whimsical ones, does not affect the agent's preferred outcomes, so that such information can be disregarded for practical purposes. But also another pattern becomes obvious, as soon as we focus our attention on the player's 'preferred option' (i.e. the actual pressure exerted by the agent's motivation on belief dynamics), represented by the first letter in each cell of Table 9. Doing so, a rather surprising fact emerges: as far as motivational pressure is concerned, it is relevant to know whether the opponent is social-oriented or not, but *it does not matter knowing the specific motivational profile which characterizes the agent's social attitude* (either Open-minded, Agreeable, Contentious or Polemist). This reduces relevant beliefs on the opponent's motivation to triadic alternative: *Conservative*, *Explorative*, or *Social-oriented* (Table 10).

Agent's profile	*Opponent is assumed to be:*		
	Conservative	*Explorative*	*Social-oriented*
Cons	***M***	***M***	***M***
Expl	***R***	***R***	***R***
Open	***R***	***M***	***M***
Agre	***R***	***M***	***R***
Cont	***M***	***R***	***M***
Pole	***M***	***R***	***R***

Table 10. Motivational pressure under different assumptions

The final output of this analysis is that an agent, to be influenced by motivation in her belief dynamics, has often to assume something on the other agent's motives, but (1) this 'something' is not everything, i.e. *exhaustive assumptions on the opponent's preferences are not needed or even useful*, and (2) it is possible, given the agent's motivational profile, *to predict what kind of motivational pressures would be produced* by each class of assumptions.

To start with the most trivial cases, *individual-oriented agents are always affected by the same motivational bias*, regardless their beliefs on the other agent's preferences: Conservative agents will always prefer to maintain previous beliefs, while Explorative ones will be more inclined to change their

introduction, and Bacharach & Bernasconi, 1997 for experimental verification). «In VFT strategies are chosen in a way which is rational in a perfectly familiar game-theoretical sense. However, the *game* that gets played is determined by nonrational (though not *ir*-rational) features of the players. These are the players' ''frames.'' A player's frame is, most simply, the set of variables she uses to conceptualize the game» (Bacharach & Bernasconi, 1997). These variables include, among other features, the players' beliefs on the opponents' preferences.

convictions. The analysis becomes much more interesting, as usual, when it comes to social-oriented attitudes. If we take for instance an Open-minded agent, we can now predict the social and motivational pressure exerted on her belief dynamics, depending on her assumptions on the other agent: if she believes her counterpart to be Conservative, she will be biased towards *revising* her previous belief; on the contrary, if she has reasons to assume the other agent to be either Explorative or Social-oriented, she will be inclined towards *confirmation* of her own views; finally, if she has no hints on the other agent's attitude, *no motivational influence is to be expected* (at least, none of the social kind). Similar considerations apply to Agreeable, Contentious and Polemist agents as well, along the lines presented in Table 10.

EXTENSION TO THE SINGLE-AGENT CASE: MOTIVATED MONOLOGUES

We are currently working on some extensions of the present framework, built on motivational attitudes other than those summarized in Table 2. For instance, while this paper was mainly focused on the *interaction* between two agents to tackle the problem of social influence in belief dynamics, we are also extending this approach to the *individual case*, i.e. a single agent who internally argues over two (possibly contradictory) claims. This requires some major adjustments in the basic assumptions, but the formal machinery remains the same. Table 11 summarizes the basic idea behind this extension: here the relevant interaction is internal to the agent's mind, involving a *pre-existing belief* and a *new incoming information* (hence the relevant axis is time, rather than sociality); the agent can (1) either maintain or drop the old belief, and (2) either accept or reject the new datum – which results in four possible outcomes. On these available options the agent can show different basic attitudes, both *past-directed* (concerning what to do with old convictions) and *future-directed* (concerning preferred reaction to new information). In turn, the interplay between past-directed and future-directed attitudes defines eight motivational profiles, which call for considerations similar to those presented so far, although leading to different conclusions.

		NEW INFORMATION	
		Accept	**Reject**
OLD BELIEF	**Maintain**	*expansion*	*stasis*
	Drop	*revision*	*contraction*

Attitude	**Criterion (partial ordering)**
Past-directed:	
- CONFIDENT	Maintain is preferred over Drop
- UNSURE	Drop is preferred over Maintain
Future-directed:	
- TRUSTFUL	Accept is preferred over Reject
- SCEPTICAL	Reject is preferred over Accept

Table 11. Extension to the individual case

Several points are worth noticing in this picture. First of all, the four possible outcomes of doxastic dynamics summarized in the upper half of Table 11 are closely related to the basic operations of AGM belief revision (Alchourrón et al., 1985; Gärdenfors, 1988): expansion, contraction and revision. Here we contemplate also the (trivial) case of stasis, in which the old belief state is maintained and the new information is rejected. This case was left aside in the original AGM paradigm, since there the incoming information was assumed to be always fully reliable, and as such accepted by the agent – a very idealistic assumption, that came to be relaxed in more recent works on non-prioritized belief revision (Hansson et al., 2001). Another point of divergence from AGM is in the interpretation of expansion and revision. According to AGM, when an agent receives a new piece of information, she performs either an expansion or a revision depending only on the semantic compatibility of the new datum with previous beliefs: if the new information does not contradict previous assumptions, we have an expansion of the belief set; otherwise, we have a revision, conceptualized as a contraction (dropping beliefs in contrast with the new datum) followed by an expansion (adding the datum to the agent belief set). The framework we are outlining in this paper is quite different in this respect: here we are not interested in the semantic value of old beliefs and new information (their contents), because we do not aim to describe the actual outcome of belief change, but rather the systematic bias that might affect such dynamics, and the underlying motivations behind that bias.

As for combining basic attitudes in complete motivational profiles, this is done along the same lines we discussed before. The eight resulting profiles are summarized in Table 12.

Past-oriented Motivational Profiles

		Primary Past-directed Attitude	
		CONF	UNSU
Secondary Future-directed Attitude	TRUS	*Self-assured* *(MA>MR>DA>DR)*	*Disproving* *(DA>DR>MA>MR)*
	SCEP	*Conservative* *(MR>MA>DR>DA)*	*Disbelieving* *(DR>DA>MR>MA)*

Future-oriented Motivational Profiles

		Primary Future-directed Attitude	
		TRUS	SCEP
Secondary Past-directed Attitude	CONF	*Novelty-seeking* *(MA>DA>MR>DR)*	*Over-cautious* *(MR>DR>MA>DA)*
	UNSU	*Flexible* *(DA>MA>DR>MR)*	*Novelty-fearing* *(DR>MR>DA>MA)*

Table 12. Motivational profiles in monological belief dynamics

These motivational profiles are modelled as *preference orderings over different operations* on the agent's belief state. For instance, an agent is conceived as Self-assured when her preferences are as follows: expansion > stasis > revision > contraction. Such an agent shows a strong bias toward confirmation of previous beliefs, together with a weaker bias towards acceptance of new information. In contrast, an agent is regarded as Disproving when a trustful attitude is coupled with (and dominated by) strong scepticism on the reliability of past beliefs: in this case, the agent preferences (revision > contraction > expansion > stasis) reveals a bias towards undermining previous convictions in favour of new suggestions. Slightly different is the case of a Disbelieving agent, with preferences contraction > revision > stasis > expansion: here systematic doubt on past beliefs is coupled with distrust towards new evidence, so that the agent is simply inclined to disbelieve most of the claims she is presented with. In a similar fashion, all other motivational profiles summarized in Table 12 can be characterized in cognitive and behavioural terms.

But what is the import of this frame of analysis for the study of individual belief dynamics? Since we are now considering the single-agent case, we cannot address the composition of the agent's preferences with those of any 'counterpart', as we did for the multi-agent case. Indeed, the very idea of an agent 'arguing with herself' is to be understood mainly as a metaphor, which does not imply any notion of a 'divided self'. On the contrary, each agent is assumed to have one and only one motivational profile at a given time.

This was actually to be expected, since, as we mentioned before, here the relevant dimension is *time*, not sociality. So an interesting set of problems to be tackled within this framework concerns the *dynamics* of these preference orderings: How do they change over time, e.g. as a function of past experiences and current context? Under which conditions an agent is expected to take on a certain motivational profile in monological belief dynamics? Is there any rationale behind such preference orderings and their dynamics?

Although detailed analysis of such issues is reserved to future works (see next section), it seems we are faced here with yet another instances of the complex interplay between motives and beliefs. In fact, while motivational profiles are likely to affect deeply the agent's belief dynamics, on the other hand they are likely to be affected in turn by the agent's understanding of past experiences – that is, by (some of) her beliefs. Apart from natural inclination, why should an agent feel either confident or unsure of past convictions? And why should she be trustful or sceptical towards new information? Clearly enough, here the agent's own assessment of past experiences and contextual features plays a crucial role in shaping her motivational profile. An initially trustful agent might easily become cautious, after being repeatedly disappointed by her informants, or when faced with unknown or disconcerting environments – and vice versa. Hence what we have here is a bidirectional pattern of influence: not only motivational profiles influence the process of belief change, but also beliefs retro-act on motives, contributing to determine the agent's preferences and their change over time.

This mutual interaction warrants further investigations, also in light of its close connections with current challenges in the field of belief revision formalisms. In our approach motivational profiles are captured as preference orderings, but such preferences do not concern beliefs or belief sets (as it is the case with most of the current belief revision models), but rather the *operations* to be performed over such sets, so that different motivational profiles come to represent *different preferential strategies for belief dynamics*. Recently (Rott, 2004) it has been argued that it is exactly at this meta-level that the role of motivations over belief change is to be investigated and understood, and several important philosophical issues are to be addressed (e.g. determinism vs. voluntarism in doxastic change). Our framework provides an operational version of some key concepts involved in this debate, offering a formal background for both theoretical analysis and experimental verification.

CONCLUSIONS AND FUTURE WORK

The game-theoretical approach presented here proved to be insightful for modelling the relation between some cognitive features of an agent (motivations and beliefs on other agents' motivations) and the specific type of social influence to be expected over her belief dynamics. This also led to specify what assumptions on the other agent's motives are truly relevant in triggering and shaping social influence, and how they interact with the agent's own preferences.

An extension of the model to the single-agent case was shortly outlined, stressing similarities and differences with the multi-agent case, and showing the relevance of this framework for current research on belief dynamics. The analysis of the single-agent case is still under development, and we are considering integrating the results summarized here in a more comprehensive model of belief change, i.e. Data-oriented Belief Revision (Paglieri, 2004).

Finally, the predictive nature of this model opens the way to empirical verification of its theoretical predictions. The direction we are currently exploring is twofold: on the one hand, we aim to test some of the model predictions in natural cognitive systems, both via economical experiments and through structural analysis of natural arguments (Paglieri & Castelfranchi, 2004); on the other hand, we are trying to implement motivational influence over belief dynamics in artificial agents, e.g. via agent-based social simulation (Castelfranchi, 1998; Paglieri, 2005).

ACKNOWLEDGMENTS

This research, as part of the ESF EUROCORES Programme *OMLL – The Origin of Man, Language and Languages*, was supported by funds from the CNR and the EC Sixth Framework Programme under Contract no. ERAS-CT-2003-980409, and by the PAR 2003 Project *Prospective Minds: Toward a theory of anticipatory mental representations*, University of Siena.

REFERENCES

Alchourrón, C., Gärdenfors P. & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic* **50**, 510-530.

Asch, S. (1952). *Social Psychology*. Englewood Cliffs: Prentice Hall.

Bacharach, M. (1993). Variable universe games. In K; Binmore et al. (Eds.), *Frontiers of Game Theory*. Cambridge: The MIT Press.

Bacharach, M., & Bernasconi, M. (1997). The variable frame theory of focal points: An experimental study. *Games and Economic Behavior* **19**, 1-45.

Berger, J. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer-Verlag.

Boutilier, C. (1998). A unified model of qualitative belief change: A dynamical systems perspective. *Artificial Intelligence* **98**, 281-316.

Castelfranchi, C. (1996). Reasons: Belief support and goal dynamics. *Mathware & Soft Computing* **3**, 233-247.

Castelfranchi, C. (1998). Modelling social action for AI agents. *Artificial Intelligence* **103**, 157-182.

Cherniak, C. (1986). *Minimal rationality*. Cambridge: The MIT Press.

Doyle, J. (1992). Reason maintenance and belief revision: Foundations vs. coherence theories. In P. Gärdenfors (Ed.), *Belief revision*. Cambridge: CUP.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford: Stanford University Press.

Forgas, J. (Ed.) (2000). *Feeling and thinking: The role of affect in social cognition*. Cambridge: CUP.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge: The MIT Press.

Goldman, A. (1999). *Knowledge in a social world*. Oxford: Clarendon Press.

Hansson, S., Fermé, E., Cantwell, J. & Falappa, M. (2001). Credibility limited revision. *Journal of Symbolic Logic* **66**, 1581-1596.

Harman, G. (1986). *Changes in view: Principles of reasoning*. Cambridge: The MIT Press.

Huns, M. & Bridgeland, D. (1991). Multi-agent Truth Maintenance. *IEEE Transactions on Systems, Man and Cybernetics* **21**, 1437-1445.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk *Econometrica* **47**, 263-291.

Kruglanski, A. (1980). Lay epistemology process and contents. *Psychological Review* **87**, 70-87.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin* **108**, 480-498.

Levi, I. (1991). *The fixation of belief and its undoing*. Cambridge: CUP.

Miceli, M. & Castelfranchi, C. (1998). Denial and its reasoning. *British Journal of Medical Psychology* **71**, 139-152.

Paglieri, F. (2004). Data-oriented Belief Revision: Toward a unified theory of epistemic processing. In E. Onaindia & S. Staab (Eds.), *Proceedings STAIRS 2004*. Amsterdam: IOS Press, pp. 179-190.

Paglieri, F. (2005). See what you want, believe what you like: Relevance and likeability in belief dynamics. In L. Cañamero (Ed.), *Proceedings AISB'05 Symposium 'Agents that want and like'*, Hatfield, pp. 90-97.

Paglieri, F. & Castelfranchi, C. (2004). Revising beliefs through arguments. In I. Rahwan, P. Moraïtis & C. Reed (Eds.), *Argumentation in MAS*. Berlin: Springer.

Picard, R. (1997). *Affective Computing*. Cambridge: The MIT Press.

Rott, H. (2004). A Counterexample to Six Fundamental Principles of Belief Formation. *Synthese* **139**, 225-240.

Swann, W. (1990). To be adored or to be known? The interplay of self-enhancement and self-verification. In R. Sorrentino & E. Higgins (Eds.), *Foundations of social behavior*. New York: Guilford.

Todd, P. & Gigerenzer, G. (2000). Simple heuristics that make us smart. *Brain and Behavioural Sciences* **23**, 727-741.

Wansing, H. (2000). A reduction of doxastic logic to action logic. *Erkenntnis* **53**, 267-283.